A short characterization of relative entropy

A short characterization of relative entropy

Tom Leinster*

School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, Scotland.

Abstract

We prove characterization theorems for relative entropy (also known as Kullback–Leibler divergence), q-logarithmic entropy (also known as Tsallis entropy), and q-logarithmic relative entropy. All three have been characterized axiomatically before, but we show that earlier proofs can be simplified considerably, at the same time relaxing some of the hypotheses.

* https://www.maths.ed.ac.uk/∼tl; Tom.Leinster@ed.ac.uk

I. INTRODUCTION

The Shannon entropy of a finite probability distribution $\mathbf{p} = (p_1, \dots, p_n)$,

$$H(\mathbf{p}) = \sum_{i:\ p_i > 0} p_i \log \frac{1}{p_i},$$

is such an important quantity that many authors have sought short lists of properties that determine H uniquely. Many such characterization theorems have been found, beginning with one in Shannon's seminal paper of 1948 (Ref. 1, Theorem 2). For instance, Faddeev [2] proved that up to a constant factor, H is uniquely characterized by symmetry, continuity, and a certain recursivity property.

Accompanying Shannon entropy is the concept of relative entropy, defined as follows. Given probability distributions $\mathbf{p}$ and $\mathbf{r}$ on n elements, the **entropy of p relative to r** is

$$H(\mathbf{p} \,\|\, \mathbf{r}) = \sum_{i:\ p_i > 0} p_i \log \frac{p_i}{r_i} \in [0, \infty].$$

Relative entropy goes by a multitude of names: Kullback–Leibler divergence, directed divergence, discrimination information, relative information, information gain, and so on. In information theory, it measures the wastage when a language whose n letters have frequencies $\mathbf{p} = (p_1, \dots, p_n)$ is encoded using a system optimized for a different language with frequencies $\mathbf{r}$, instead of the system optimized for the original language. There are other interpretations in other fields, as the plethora of names suggests.

Axiomatic characterizations of relative entropy have also been sought and found. One such theorem is implicit in work of Kannappan and Ng [3]. It states that up to a constant factor, relative entropy is uniquely determined by measurability in each of $\mathbf{p}$ and $\mathbf{r}$ separately, invariance under permutations of $\{1, \dots, n\}$, the vanishing property $H(\mathbf{p} \,\|\, \mathbf{p}) = 0$, and a certain recursivity equation. (Remark II.7 gives further details.) Their proof was a tour de force of functional equations, involving the solution of the functional equation

$$f(x) + (1-x)g\left(\frac{y}{1-x}\right) = h(y) + (1-y)k\left(\frac{x}{1-y}\right) \tag{1}$$

in four unknown functions, as well as the four-variable functional equation

$$F(x,y) + (1-x)F\left(\frac{u}{1-x}, \frac{v}{1-y}\right) = F(u,v) + (1-u)F\left(\frac{x}{1-u}, \frac{y}{1-v}\right).$$

We give a much simpler proof, at the same time weakening the measurability hypothesis. Our proof involves neither of these equations. Instead, it borrows heavily from a categorical characterization of relative entropy by Baez and Fritz [4], which in turn was inspired by a characterization by Petz [5] of the relative entropy of states of matrix algebras. Our characterization of relative entropy is the first main result, Theorem II.1.

Shannon entropy is just one member (albeit a special one) of a one-parameter family of entropies $(S_q)_{q\in\mathbb{R}}$, first investigated by Havrda and Charvát [6] and often misattributed to Tsallis (Remark III.2(ii)). These entropies S_q, and the accompanying relative entropies, are defined as follows.

For $q \in \mathbb{R}$, the **q-logarithm** is the function $\ln_q \colon (0,\infty) \to \mathbb{R}$ given by

$$\ln_q(x) = \int_1^x t^{-q}\, dt.$$

The **q-logarithmic entropy** and **q-logarithmic relative entropy** are defined by

$$S_q(\mathbf{p}) = \sum_{i\colon p_i>0} p_i \ln_q \frac{1}{p_i},$$
$$S_q(\mathbf{p} \,\|\, \mathbf{r}) = -\sum_{i\colon p_i>0} p_i \ln_q \frac{r_i}{p_i},$$

for probability distributions $\mathbf{p}$ and $\mathbf{r}$ on n elements. When $q = 1$, these reduce to the ordinary Shannon entropy and relative entropy.

There are several existing theorems characterizing the q-logarithmic entropy for a given $q \neq 1$. Up until now, the simplest appears to have been the 1970 result of Daróczy [7]. We simplify further, weakening the hypotheses and shortening the proof to just a few lines (Theorem III.1). Finally, we use a similar and equally short argument to characterize the q-logarithmic relative entropies $S_q(- \,\|\, -)$ (Theorem IV.1).

It is remarkable that when $q \neq 1$, the characterizations of q-logarithmic entropy and q-logarithmic relative entropy need no regularity conditions whatsoever (not even measurability), in contrast to the theorems for $q = 1$.

The remaining three sections of this paper establish our three theorems in turn, characterizing first relative entropy, then q-logarithmic entropy, then q-logarithmic relative entropy.

II. RELATIVE ENTROPY

For $n \geq 1$, write

$$\Delta_n = \Big\{ \mathbf{p} = (p_1, \ldots, p_n) \in \mathbb{R}^n : p_i \geq 0, \ \sum p_i = 1 \Big\}$$

for the set of probability distributions on $\{1, \ldots, n\}$, and write

$$A_n = \big\{ (\mathbf{p}, \mathbf{r}) \in \Delta_n \times \Delta_n : p_i = 0 \text{ whenever } r_i = 0 \big\}.$$

Evidently, $(\mathbf{p}, \mathbf{r}) \in A_n$ if and only if the relative entropy

$$H_n(\mathbf{p} \,\|\, \mathbf{r}) = \sum_{i:\ p_i > 0} p_i \log \frac{p_i}{r_i}$$

is finite. (In this section, we add an 'n' subscript to H for clarity.) Viewing $\mathbf{p}$ and $\mathbf{r}$ as measures on $\{1, \ldots, n\}$, we have $(\mathbf{p}, \mathbf{r}) \in A_n$ just when $\mathbf{p}$ is absolutely continuous with respect to $\mathbf{r}$.

We will characterize the sequence of functions

$$H(- \,\|\, -) = \big(H_n(- \,\|\, -) \colon A_n \to \mathbb{R} \big)_{n \geq 1}$$

uniquely up to a constant factor. It is easy to check that this sequence has the following four properties, as does any scalar multiple $cH(- \,\|\, -)$ with $c \in \mathbb{R}$.

Measurability in the second argument: For each $n \geq 1$ and $\mathbf{p} \in \Delta_n$, the function

$$\begin{array}{rcl} \{\mathbf{r} \in \Delta_n : (\mathbf{p}, \mathbf{r}) \in A_n\} & \to & \mathbb{R} \\ \mathbf{r} & \mapsto & H_n(\mathbf{p} \,\|\, \mathbf{r}) \end{array}$$

is Lebesgue measurable.

Symmetry: For each $n \geq 1$, $(\mathbf{p}, \mathbf{r}) \in A_n$ and permutation σ of $\{1, \ldots, n\}$,

$$H_n(\mathbf{p} \,\|\, \mathbf{r}) = H_n(\mathbf{p}\sigma \,\|\, \mathbf{r}\sigma), \tag{2}$$

where $\mathbf{p}\sigma = (p_{\sigma(1)}, \ldots, p_{\sigma(n)})$.

Vanishing: $H_n(\mathbf{p} \,\|\, \mathbf{p}) = 0$ for all $n \geq 1$ and $\mathbf{p} \in \Delta_n$.

Chain rule: To state this, we need some notation. Given $n, k_1, \ldots, k_n \geq 1$ and $\mathbf{w} \in \Delta_n, \mathbf{p}^1 \in \Delta_{k_1}, \ldots, \mathbf{p}^n \in \Delta_{k_n}$, and writing $\mathbf{p}^i = (p^i_1, \ldots, p^i_{k_i})$, define

$$\mathbf{w} \circ (\mathbf{p}^1, \ldots, \mathbf{p}^n) = (w_1 p^1_1, \ldots, w_1 p^1_{k_1}, \ \ldots, \ w_n p^n_1, \ldots, w_n p^n_{k_n}) \in \Delta_{k_1 + \cdots + k_n}.$$

The **chain rule** for relative entropy is that

$$H_{k_1+\cdots+k_n}\big(\mathbf{w}\circ(\mathbf{p}^1,\dots,\mathbf{p}^n)\,\big\|\,\widetilde{\mathbf{w}}\circ(\widetilde{\mathbf{p}}^1,\dots,\widetilde{\mathbf{p}}^n)\big) = H_n(\mathbf{w}\,\|\,\widetilde{\mathbf{w}}) + \sum_{i=1}^n w_i H_{k_i}(\mathbf{p}^i\,\|\,\widetilde{\mathbf{p}}^i) \qquad (3)$$

whenever $(\mathbf{w},\widetilde{\mathbf{w}})\in A_n$ and $(\mathbf{p}^i,\widetilde{\mathbf{p}}^i)\in A_{k_i}$. (Under these hypotheses, the pair of distributions on the left-hand side belongs to $A_{k_1+\cdots+k_n}$.)

Theorem II.1 *Let $I(-\,\|\,-) = \big(I_n(-\,\|\,-)\colon A_n\to\mathbb{R}\big)_{n\geq 1}$ be a sequence of functions. The following are equivalent:*

i. $I(-\,\|\,-)$ satisfies the four properties above: measurability in the second argument, symmetry, vanishing, and the chain rule;

ii. $I(-\,\|\,-) = cH(-\,\|\,-)$ for some $c\in\mathbb{R}$.

We have just noted that (ii) implies (i). We now embark on the proof of the converse. For the rest of this section, let $I(-\,\|\,-) = \big(I_n(-\,\|\,-)\colon A_n\to\mathbb{R}\big)_{n\geq 1}$ be a sequence of functions satisfying the four conditions. Define a function $L\colon (0,1]\to\mathbb{R}$ by

$$L(\alpha) = I_2\big((1,0)\,\big\|\,(\alpha,1-\alpha)\big).$$

The idea is that if $I(-\,\|\,-) = H(-\,\|\,-)$ then $L = -\log$. We will show that in any case, L is a scalar multiple of log.

Lemma II.2 *Let $(\mathbf{p},\mathbf{r})\in A_n$ with $p_{k+1} = \cdots = p_n = 0$, where $1\leq k\leq n$. Then $r_1+\cdots+r_k > 0$ and*

$$I_n(\mathbf{p}\,\|\,\mathbf{r}) = L(r_1+\cdots+r_k) + I_k(\mathbf{p}'\,\|\,\mathbf{r}'),$$

where

$$\mathbf{p}' = (p_1,\dots,p_k), \qquad \mathbf{r}' = \frac{(r_1,\dots,r_k)}{r_1+\cdots+r_k}.$$

Proof The case $k=n$ reduces to the statement that $L(1)=0$, which follows from the vanishing property. Suppose, then, that $k<n$.

Since $\mathbf{p}$ is a probability distribution with $p_i = 0$ for all $i>k$, there is some $i\leq k$ such that $p_i>0$, and then $r_i>0$ as $(\mathbf{p},\mathbf{r})\in A_n$. Hence $r_1+\cdots+r_k>0$. Let $\mathbf{r}''\in\Delta_{n-k}$ be the normalization of $(r_{k+1},\dots,r_n)$ if $r_{k+1}+\cdots+r_n>0$, or choose $\mathbf{r}''$ arbitrarily in Δ_{n-k} otherwise (which is possible since $k<n$). Then

$$I_n(\mathbf{p}\,\|\,\mathbf{r}) = I_n\big((1,0)\circ(\mathbf{p}',\mathbf{r}'')\,\big\|\,(r_1+\cdots+r_k,\, r_{k+1}+\cdots+r_n)\circ(\mathbf{r}',\mathbf{r}'')\big).$$

The result now follows from the chain rule. □

Lemma II.3 *$L(\alpha\beta) = L(\alpha) + L(\beta)$ for all $\alpha, \beta \in (0,1]$.*

Proof We evaluate the real number

$$x := I_3\big((1,0,0) \,\big\|\, (\alpha\beta, \alpha(1-\beta), 1-\alpha)\big)$$

in two ways. By Lemma II.2 with $k = 1$ and the vanishing property,

$$x = L(\alpha\beta) + I_1\big((1) \,\big\|\, (1)\big) = L(\alpha\beta),$$

where (1) denotes the unique element of Δ_1. But also, by Lemma II.2 with $k = 2$,

$$x = L(\alpha) + I_2\big((1,0) \,\big\|\, (\beta, 1-\beta)\big) = L(\alpha) + L(\beta).$$

Comparing the two expressions for x gives the result. □

Lemma II.4 *There is some $c \in \mathbb{R}$ such that $L(\alpha) = -c \log \alpha$ for all $\alpha \in (0,1]$.*

Proof Define $f\colon [0,\infty) \to \mathbb{R}$ by $f(t) = L(e^{-t})$. By Lemma II.3, f satisfies Cauchy's functional equation: $f(t+u) = f(t) + f(u)$ for all $t, u \in [0,\infty)$. Also, f is measurable, since L is. It is well-known [8] that these conditions force $f(t) = ct$ for some constant c, giving $L(\alpha) = -c \log \alpha$. □

Our next lemma is an adaptation of the most ingenious part of Baez and Fritz's argument (Ref. 4, Lemma 4.2).

Lemma II.5 *Let $(\mathbf{p}, \mathbf{r}) \in A_n$ with $p_i > 0$ for all $i \in \{1, \dots, n\}$. Then $I_n(\mathbf{p} \,\|\, \mathbf{r}) = cH_n(\mathbf{p} \,\|\, \mathbf{r})$.*

Proof The hypotheses imply that $r_i > 0$ for all i. We can therefore choose some $\alpha \in (0,1]$ such that $r_i - \alpha p_i \geq 0$ for all i. We will compute the number

$$x := I_{2n}\big((p_1, \dots, p_n, \underbrace{0, \dots, 0}_{n}) \,\big\|\, (\alpha p_1, \dots, \alpha p_n, r_1 - \alpha p_1, \dots, r_n - \alpha p_n)\big)$$

in two ways. First, by Lemma II.2, Lemma II.4, and the vanishing property,

$$x = L(\alpha) + I_n(\mathbf{p} \,\|\, \mathbf{p}) = -c \log \alpha.$$

Second, by symmetry, the chain rule, and Lemma II.4,

$$\begin{aligned} x &= I_{2n}\big((p_1, 0, \dots, p_n, 0) \,\big\|\, (\alpha p_1, r_1 - \alpha p_1, \dots, p_n, r_n - \alpha p_n)\big) \\ &= I_{2n}\Big(\mathbf{p} \circ \big((1,0), \dots, (1,0)\big) \,\Big\|\, \mathbf{r} \circ \big(\big(\alpha \tfrac{p_1}{r_1}, 1 - \alpha \tfrac{p_1}{r_1}\big), \dots, \big(\alpha \tfrac{p_n}{r_n}, 1 - \alpha \tfrac{p_n}{r_n}\big)\big)\Big) \\ &= I_n(\mathbf{p} \,\|\, \mathbf{r}) + \sum_{i=1}^{n} p_i L\big(\alpha \tfrac{p_i}{r_i}\big) \\ &= I_n(\mathbf{p} \,\|\, \mathbf{r}) - c \log \alpha - cH_n(\mathbf{p} \,\|\, \mathbf{r}). \end{aligned}$$

Comparing the two expressions for x gives the result. □

We have now proved that $I_n(\mathbf{p} \,\|\, \mathbf{r}) = cH_n(\mathbf{p} \,\|\, \mathbf{r})$ when $\mathbf{p}$ has full support. It only remains to prove it for arbitrary $\mathbf{p}$.

Proof of Theorem II.1 Let $(\mathbf{p}, \mathbf{r}) \in A_n$. By symmetry, we can assume that $p_1, \dots, p_k > 0$ and $p_{k+1} = \cdots = p_n = 0$ for some $k \in \{1, \dots, n\}$. Writing $R = r_1 + \cdots + r_k$, we have $R > 0$ since $(\mathbf{p}, \mathbf{r}) \in A_n$, and then

$$I_n(\mathbf{p} \,\|\, \mathbf{r}) = L(R) + I_k\big((p_1, \dots, p_k) \,\big\|\, \tfrac{1}{R}(r_1, \dots, r_k)\big)$$

by Lemma II.2. Hence by Lemmas II.4 and II.5,

$$I_n(\mathbf{p} \,\|\, \mathbf{r}) = -c \log R + cH_k\big((p_1, \dots, p_k) \,\big\|\, \tfrac{1}{R}(r_1, \dots, r_k)\big).$$

But by the same argument applied to cH in place of I (or by direct calculation), we also have

$$cH_n(\mathbf{p} \,\|\, \mathbf{r}) = -c \log R + cH_k\big((p_1, \dots, p_k) \,\big\|\, \tfrac{1}{R}(r_1, \dots, r_k)\big).$$

The result follows. □

Remarks II.6 i. The vanishing axiom cannot be dropped from Theorem II.1. Indeed, the quantity $\sum_{i:\ p_i > 0} p_i \log \frac{1}{r_i}$ satisfies the other three axioms but not vanishing.

ii. In the literature on information functions, the chain rule is often replaced by one of two superficially simpler rules. The first is the special case $k_1 = 2, k_2 = \cdots = k_n = 1$, which is

$$\begin{aligned} H_{n+1}\big((pw_1, (1-p)w_1, w_2, \dots, w_n) \,\big\|\, (\widetilde{p}\widetilde{w}_1, (1-\widetilde{p})\widetilde{w}_1, \widetilde{w}_2, \dots, \widetilde{w}_n)\big) \\ = H_n\big(\mathbf{w} \,\big\|\, \widetilde{\mathbf{w}}\big) + w_1 H_2\big((p, 1-p) \,\big\|\, (\widetilde{p}, 1-\widetilde{p})\big) \quad (4) \end{aligned}$$

($(\mathbf{w}, \widetilde{\mathbf{w}}) \in A_n$, $((p, 1-p), (\widetilde{p}, 1-\widetilde{p})) \in A_2$). This is known as **recursivity** or **grouping**. The second is the special case $n = 2$ of the chain rule, which is

$$\begin{aligned} H_{k+\ell}\big(w\mathbf{p} \oplus (1-w)\mathbf{r} \,\big\|\, \widetilde{w}\widetilde{\mathbf{p}} \oplus (1-\widetilde{w})\widetilde{\mathbf{r}}\big) \\ = H_2\big((w, 1-w) \,\big\|\, (\widetilde{w}, 1-\widetilde{w})\big) + wH_k(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) + (1-w)H_\ell(\mathbf{r} \,\|\, \widetilde{\mathbf{r}}), \quad (5) \end{aligned}$$

where

$$w\mathbf{p} \oplus (1-w)\mathbf{r} = (wp_1, \dots, wp_k, (1-w)r_1, \dots, (1-w)r_\ell)$$

and $((w, 1-w), (\widetilde{w}, 1-\widetilde{w})) \in A_2$, $(\mathbf{p}, \widetilde{\mathbf{p}}) \in A_k$, $(\mathbf{r}, \widetilde{\mathbf{r}}) \in A_\ell$. However, straightforward inductions show that in the presence of the symmetry axiom, either one of the special cases (4) or (5) is equivalent to the full chain rule (3). (Similar inductions are carried out in Ref. 9, p. 5–6.) Which to use is, therefore, simply a matter of taste.

Remark II.7 Here we compare Theorem II.1 with some earlier characterizations of relative entropy. One of the first such theorems was that of Hobson [10], who used stronger hypotheses for the same conclusion. In common with Theorem II.1, he assumed symmetry, vanishing, and the chain rule (in the equivalent form (5)). But he also assumed continuity in both variables (instead of measurability in one) and a monotonicity hypothesis unlike anything in Theorem II.1.

In 1973, Kannappan and Ng [3] proved a result very close to Theorem II.1. They did not *state* that result in Ref. 3, but the closing remarks in another paper by the same authors [11] and the approach of a contemporaneous paper by Kannappan and Rathie [12] strongly suggest the intent. The result was stated explicitly by Csiszár (Ref. 13, Section 2.1), who attributed it to Kannappan and Ng.

There are some slight differences of hypotheses between Kannappan and Ng's theorem and Theorem II.1. They assumed measurability in both variables, whereas we only assumed measurability in the second. (In fact, all we used was that $I_2((1,0) \,\|\, -)$ is measurable.) On the other hand, they only needed the vanishing condition for $(1/2, 1/2)$, whereas we needed it for all $\mathbf{p}$. They used the chain rule in the equivalent form (4). Their proof and ours are entirely different.

III. q-LOGARITHMIC ENTROPY

Let $q \in \mathbb{R}$. The definition of q-logarithm in the Introduction gives, explicitly,

$$\ln_q(x) = \frac{1}{1-q}(x^{1-q} - 1)$$

for $x \in (0, \infty)$ and $q \neq 1$, while $\ln_1$ is the natural logarithm log. Hence, explicitly, the q-logarithmic entropy is given by

$$S_q(\mathbf{p}) = \frac{1}{1-q}\left(\sum_{i:\, p_i > 0} p_i^q - 1\right)$$

for $\mathbf{p} \in \Delta_n$ and $q \neq 1$, while S_1 is the Shannon entropy H. We have $\ln_q(x) \to \log(x)$ as $q \to 1$, hence also $S_q(\mathbf{p}) \to H(\mathbf{p})$ as $q \to 1$.

Fix $q \in \mathbb{R}$. The q-logarithmic entropy satisfies a chain rule

$$S_q\big(\mathbf{w} \circ (\mathbf{p}^1, \ldots, \mathbf{p}^n)\big) = S_q(\mathbf{w}) + \sum_{i:\, w_i > 0} w_i^q S_q(\mathbf{p}^i) \tag{6}$$

($\mathbf{w} \in \Delta_n$, $\mathbf{p}^1 \in \Delta_{k_1}, \ldots, \mathbf{p}^n \in \Delta_{k_n}$), as is easily checked. In particular, this holds when $\mathbf{p}^1 = \cdots =$

$\mathbf{p}^n = \mathbf{p}$, say. For $\mathbf{w} \in \Delta_n$ and $\mathbf{p} \in \Delta_k$, write

$$\begin{aligned}\mathbf{w} \otimes \mathbf{p} &= \mathbf{w} \circ (\mathbf{p}, \dots, \mathbf{p}) \\ &= (w_1 p_1, \dots, w_1 p_k, \ \dots, \ w_n p_1, \dots, w_n p_k) \in \Delta_{nk}.\end{aligned}$$

In this case, the q-chain rule (6) gives a **q-multiplicativity** property:

$$S_q(\mathbf{w} \otimes \mathbf{p}) = S_q(\mathbf{w}) + \left(\sum_{i:\, w_i > 0} w_i^q \right) S_q(\mathbf{p}) \tag{7}$$

($n, k \geq 1$, $\mathbf{w} \in \Delta_n$, $\mathbf{p} \in \Delta_k$).

Note also that S_q is symmetric in its arguments:

$$S_q(\mathbf{p}) = S_q(\mathbf{p}\sigma) \tag{8}$$

for all $\mathbf{p} \in \Delta_n$ and permutations σ of $\{1, \dots, n\}$.

The left-hand side of equation (7) is symmetric in $\mathbf{w}$ and $\mathbf{p}$, but the right-hand side is not obviously so. This is the key to our second theorem.

Theorem III.1 *Let $1 \neq q \in \mathbb{R}$ and let $I = (I_n \colon \Delta_n \to \mathbb{R})_{n \geq 1}$ be a sequence of functions. The following are equivalent:*

i. I has the q-multiplicativity property (7) *and the symmetry property* (8) *(both with I in place of S_q);*

ii. $I = cS_q$ for some $c \in \mathbb{R}$.

Proof By the observations just made, (ii) implies (i). Now assume (i). For all $\mathbf{w} \in \Delta_n$ and $\mathbf{p} \in \Delta_k$, we have $I_{nk}(\mathbf{w} \otimes \mathbf{p}) = I_{nk}(\mathbf{p} \otimes \mathbf{w})$ by symmetry, so

$$I_n(\mathbf{w}) + \left(\sum_{i:\, w_i > 0} w_i^q \right) I_k(\mathbf{p}) = I_k(\mathbf{p}) + \left(\sum_{i:\, p_i > 0} p_i^q \right) I_n(\mathbf{w}),$$

or equivalently

$$\left(\sum_{i:\, w_i > 0} w_i^q - 1 \right) I_k(\mathbf{p}) = \left(\sum_{i:\, p_i > 0} p_i^q - 1 \right) I_n(\mathbf{w}).$$

Take $\mathbf{w} = (1/2, 1/2)$: then for all $\mathbf{p} \in \Delta_k$,

$$\left(2^{1-q} - 1\right) I_k(\mathbf{p}) = \left(\sum_{i:\, p_i > 0} p_i^q - 1 \right) I_2(1/2, 1/2).$$

Since $q \neq 1$, we can define $c = \frac{1-q}{2^{1-q}-1} \cdot I_2(1/2, 1/2)$, and then $I = cS_q$. □

Remarks III.2 i. The q-logarithms were used in Hardy, Littlewood and Pólya's classic book on inequalities, first published in 1934 (Ref. 14, proof of Theorem 84). They have been an explicit object of study since at least a 1964 paper of Box and Cox in statistics (Ref. 15, Section 3). The name 'q-logarithm' appears to have been introduced by Umarov, Tsallis and Steinberg [16] in 2008, working in statistical mechanics.

ii. The q-logarithmic entropies have been discovered and rediscovered repeatedly. They seem to have first appeared in a 1967 paper on information and classification by Havrda and Charvát [6], who used a form adapted to base 2 logarithms. They were rediscovered in 1970 by Daróczy [7]. The base e version S_q appeared in a 1982 article of Patil and Taillie (Ref. 17, Section 3.2), where it was studied as an index of biodiversity.

In physics, meanwhile, the q-logarithmic entropies appeared in a 1971 article of Lindhard and Nielsen [18] (according to Csiszár: Ref. 13, Section 2.4), and in a 1978 survey by Wehrl (Ref. 19, p. 247). Finally, they were rediscovered again in a 1988 paper on statistical physics by Tsallis [20].

Despite the twenty years of active life that the q-logarithmic entropies had already enjoyed, it is after Tsallis that they are most commonly named. The term 'q-logarithmic entropy' is new, but has the benefits of being descriptive and of not perpetuating a misattribution.

iii. As in Remark II.6(ii), a simple inductive argument shows that the q-chain rule of equation (6) follows from the special case

$$S_q\big(pw_1, (1-p)w_1, w_2, \ldots, w_n\big) = S_q(\mathbf{w}) + w_1^q S_q(p, 1-p) \tag{9}$$

($p \in [0,1]$, $n \geq 1$, $\mathbf{w} \in \Delta_n$).

iv. A characterization of the q-logarithmic entropies similar to Theorem III.1 was published by Daróczy in 1970 [7]. He assumed the full q-chain rule for $I(\mathbf{w} \circ (\mathbf{p}^1, \ldots, \mathbf{p}^n))$ (in the equivalent form (9)), rather than just the special case $\mathbf{p}^1 = \cdots = \mathbf{p}^n$ used here. However, where we assumed that $I_n \colon \Delta_n \to \mathbb{R}$ is symmetric for all $n \geq 2$, Daróczy only assumed it for $n = 3$. The two proofs are very different; the main step in Daróczy's was the solution of the functional equation (1) in the case $f = g = h = k$.

Other characterizations of S_q have been proved, using stronger hypotheses than Theorem III.1 to obtain the same conclusion (such as the theorem in Section 2 of Ref. 21, and Theorem V.2 of Ref. 22).

IV. q-LOGARITHMIC RELATIVE ENTROPY

For $q \neq 1$, the q-logarithmic relative entropy $S_q \colon A_n \to \mathbb{R}$, defined in the Introduction, is given explicitly by

$$S_q(\mathbf{p} \,\|\, \mathbf{r}) = \frac{1}{q-1}\left(\sum_{i\colon p_i > 0} p_i^q r_i^{1-q} - 1\right)$$

for $(\mathbf{p}, \mathbf{r}) \in A_n$. In the case $q = 1$, it reduces to the ordinary relative entropy $H(\mathbf{p} \,\|\, \mathbf{r})$. As in that case, restricting the arguments to lie in A_n guarantees that $S_q(- \,\|\, -)$ takes only finite values.

Our third and final theorem is a characterization of q-logarithmic relative entropy, very similar to the characterization of q-logarithmic entropy itself.

We begin by noting two properties of q-logarithmic relative entropy. First, there is an easily-checked chain rule:

$$S_q\big(\mathbf{w} \circ (\mathbf{p}^1, \ldots, \mathbf{p}^n) \,\big\|\, \widetilde{\mathbf{w}} \circ (\widetilde{\mathbf{p}}^1, \ldots, \widetilde{\mathbf{p}}^n)\big) = S_q(\mathbf{w} \,\|\, \widetilde{\mathbf{w}}) + \sum_{i\colon w_i > 0} w_i^q \widetilde{w}_i^{1-q} S_q(\mathbf{p}^i \,\|\, \widetilde{\mathbf{p}}^i)$$

($(\mathbf{w}, \widetilde{\mathbf{w}}) \in A_n$, $(\mathbf{p}^i, \widetilde{\mathbf{p}}^i) \in A_{k_i}$). This specializes to a q-multiplicativity formula

$$S_q(\mathbf{w} \otimes \mathbf{p} \,\|\, \widetilde{\mathbf{w}} \otimes \widetilde{\mathbf{p}}) = S_q(\mathbf{w} \,\|\, \widetilde{\mathbf{w}}) + \left(\sum_{i\colon w_i > 0} w_i^q \widetilde{w}_i^{1-q}\right) S_q(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) \tag{10}$$

($(\mathbf{w}, \widetilde{\mathbf{w}}) \in A_n$, $(\mathbf{p}, \widetilde{\mathbf{p}}) \in A_k$). Second, q-logarithmic relative entropy has the same symmetry property as ordinary relative entropy:

$$S_q(\mathbf{p} \,\|\, \mathbf{r}) = S_q(\mathbf{p}\sigma \,\|\, \mathbf{r}\sigma) \tag{11}$$

for all $n \geq 1$, $(\mathbf{p}, \mathbf{r}) \in A_n$, and permutations σ of $\{1, \ldots, n\}$.

Theorem IV.1 *Let $1 \neq q \in \mathbb{R}$ and let $I(- \,\|\, -) = \big(I_n(- \,\|\, -) \colon A_n \to \mathbb{R}\big)_{n \geq 1}$ be a sequence of functions. The following are equivalent:*

i. $I(- \,\|\, -)$ has the q-multiplicativity property (10) *and the symmetry property* (11) *(both with I in place of S_q);*

ii. $I(- \,\|\, -) = cS_q(- \,\|\, -)$ for some $c \in \mathbb{R}$.

Proof It is trivial that (ii) implies (i). Now assume (i). By symmetry,

$$I_{nk}(\mathbf{w} \otimes \mathbf{p} \,\|\, \widetilde{\mathbf{w}} \otimes \widetilde{\mathbf{p}}) = I_{nk}(\mathbf{p} \otimes \mathbf{w} \,\|\, \widetilde{\mathbf{p}} \otimes \widetilde{\mathbf{w}})$$

for all $n, k \geq 1$, $(\mathbf{w}, \widetilde{\mathbf{w}}) \in A_n$, and $(\mathbf{p}, \widetilde{\mathbf{p}}) \in A_k$. So by q-multiplicativity,

$$I_n(\mathbf{w} \,\|\, \widetilde{\mathbf{w}}) + \left(\sum_{i \colon w_i > 0} w_i^q \widetilde{w}_i^{1-q} \right) I_k(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) = I_k(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) + \left(\sum_{i \colon p_i > 0} p_i^q \widetilde{p}_i^{1-q} \right) I_n(\mathbf{w} \,\|\, \widetilde{\mathbf{w}}),$$

or equivalently,

$$\left(\sum_{i \colon w_i > 0} w_i^q \widetilde{w}_i^{1-q} - 1 \right) I_k(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) = \left(\sum_{i \colon p_i > 0} p_i^q \widetilde{p}_i^{1-q} - 1 \right) I_n(\mathbf{w} \,\|\, \widetilde{\mathbf{w}}).$$

Take $\mathbf{w} = (1, 0)$ and $\widetilde{\mathbf{w}} = (1/2, 1/2)$: then

$$(2^{q-1} - 1) I_k(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) = \left(\sum_{i \colon p_i > 0} p_i^q \widetilde{p}_i^{1-q} - 1 \right) I_2\big((1, 0) \,\big\|\, (1/2, 1/2)\big)$$

for all $(\mathbf{p}, \widetilde{\mathbf{p}}) \in A_k$. But $q \neq 1$, so we can define

$$c = \frac{1 - q}{2^{q-1} - 1} \cdot I_2\big((1, 0) \,\big\|\, (1/2, 1/2)\big),$$

and then $I(- \,\|\, -) = c S_q(- \,\|\, -)$. □

Remarks IV.2 i. The definition of q-logarithmic relative entropy was given in 1972 by Rathie and Kannappan [23] (who used a version adapted to base 2 logarithms). The base e version used here was studied by Cressie and Read in a 1984 paper in statistics (Ref. 24, Section 5). It was rediscovered in physics in 1998, by Shiino [25] and Tsallis [26] independently.

ii. Other characterization theorems for q-logarithmic relative entropy have been proved. For example, Furuichi (Ref. 22, Section IV) obtained the same conclusion, but also assumed continuity and essentially the full chain rule (that is, an equivalent special case, as in Remarks II.6(ii) and III.2(iii)).

Acknowledgements I thank John Baez and Tobias Fritz for their comments. This work was partially supported by a BBSRC FLIP award (BB/P004210/1).

---

[1] C. E. Shannon, Bell System Technical Journal **27**, 379 (1948).

[2] D. K. Faddeev, Uspekhi Matematicheskikh Nauk **11**, 227 (1956).

[3] P. Kannappan and C. T. Ng, Proceedings of the American Mathematical Society **38**, 303 (1973).

[4] J. Baez and T. Fritz, Theory and Applications of Categories **29**, 421 (2014).

[5] D. Petz, Acta Mathematica Hungarica **59**, 449 (1992).

[6] J. Havrda and F. Charvát, Kybernetika **3**, 30 (1967).

[7] Z. Daróczy, Information and Control **16**, 36 (1970).

[8] S. Banach, Fundamenta Mathematicae **1**, 123 (1920).

[9] A. Feinstein, *Foundations of Information Theory* (McGraw–Hill, New York, 1958).

[10] A. Hobson, Journal of Statistical Physics **1**, 383 (1969).

[11] P. Kannappan and C. T. Ng, Pacific Journal of Mathematics **54**, 157 (1974).

[12] P. Kannappan and P. N. Rathie, Information and Control **22**, 163 (1973).

[13] I. Csiszár, Entropy **10**, 261 (2008).

[14] G. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed. (Cambridge University Press, Cambridge, 1952).

[15] G. E. P. Box and D. R. Cox, Journal of the Royal Statistical Society. Series B (Methodological) **26**, 211 (1964).

[16] S. Umarov, C. Tsallis, and S. Steinberg, Milan Journal of Mathematics **76**, 307 (2008).

[17] G. P. Patil and C. Taillie, Journal of the American Statistical Association **77**, 548 (1982).

[18] J. Lindhard and V. Nielsen, Matematisk-Fysiske Meddelelser: Kongelige Danske Videnskabernes Selskab **38**, 1 (1971).

[19] A. Wehrl, Reviews of Modern Physics **50**, 221 (1978).

[20] C. Tsallis, Journal of Statistical Physics **52**, 479 (1988).

[21] H. Suyari, Journal of Physics A: Mathematical and General **35**, 10731 (2002).

[22] S. Furuichi, IEEE Transactions on Information Theory **51**, 3638 (2005).

[23] P. N. Rathie and P. Kannappan, Information and Control **20**, 38 (1972).

[24] N. Cressie and T. R. C. Read, Journal of the Royal Statistical Society. Series B (Methodological) **46**, 440 (1984).

[25] M. Shiino, Journal of the Physical Society of Japan **67**, 3658 (1998).

[26] C. Tsallis, Physical Review E **58**, 1442 (1998).